

Chapter 7

Perception

Perception is the foundational gateway through which both humans and intelligent agents acquire information, interpret their surroundings, and ultimately make informed decisions. For humans, perception is seamless and intuitive, effortlessly transforming sensory inputs into meaningful interpretations. In artificial intelligence, however, perception systems are meticulously engineered to emulate—and in some respects surpass—human sensory processing, profoundly influencing an agent’s capacity for interaction, learning, and adaptation in complex environments.

In this chapter, we begin by exploring key differences in the nature and efficiency of perception between humans and AI agents. Next, we categorize agent perception based on different forms and representations of perceptual input. We then discuss ongoing challenges in the agent perception system and highlight promising directions for improvement, both at the modeling and system architecture levels. Finally, we illustrate how perception modules can be effectively tailored to different intelligent agent scenarios, offering practical guidance for optimizing their use and suggesting pivotal areas for future research.

7.1 Human versus AI Perception

Perception is fundamental to intelligence, serving as the interface through which both humans and artificial agents interact with the world. Although humans commonly think of perception in terms of the five classical senses—vision, hearing, taste, smell, and touch—modern neuroscience identifies a richer sensory landscape. Conservatively, humans are described as having around 10 senses; more comprehensive views list approximately 21, while some researchers propose up to 33 distinct sensory modalities [546, 547]. Beyond the familiar senses, humans possess sophisticated internal perceptions, such as vestibular (balance), proprioception (awareness of body position), thermoception (temperature), and nociception (pain), enabling nuanced interaction with their environment.

Human senses are finely tuned to specific physical signals: for example, human vision detects electromagnetic waves with wavelengths between approximately 380–780 nm, whereas hearing perceives sound frequencies from about 20 Hz to 20 kHz [548]. These sensory modalities allow humans to effortlessly engage in complex tasks like language communication, object recognition, social interaction, and spatial navigation. Additionally, humans naturally perceive continuous changes over time, seamlessly integrating motion perception and temporal awareness, abilities essential for coordinated movement and decision-making [549]. Animals in the natural world exhibit even more diverse perceptual capabilities. Birds and certain marine organisms, for instance, utilize magnetoreception to navigate using Earth’s magnetic fields, while sharks and electric eels exploit electroreception to sense electrical signals emitted by other organisms—abilities humans do not possess [550].

In contrast to biological perception, artificial agents rely upon engineered sensors designed to transform environmental stimuli into digital signals that algorithms can interpret. Common sensor modalities for AI agents include visual sensors (cameras), auditory sensors (microphones), tactile sensors, and inertial measurement units. AI agents typically excel at processing visual, auditory, and textual data, leveraging advances in deep learning and signal processing. However, certain human sensory abilities—particularly taste and smell—remain challenging for machines to emulate accurately. For example, the advanced bio-inspired olfactory chip developed by researchers [551] currently distinguishes around 24 different odors, a capability significantly less sensitive than the human olfactory system, which discriminates among more than 4,000 distinct smells [552].

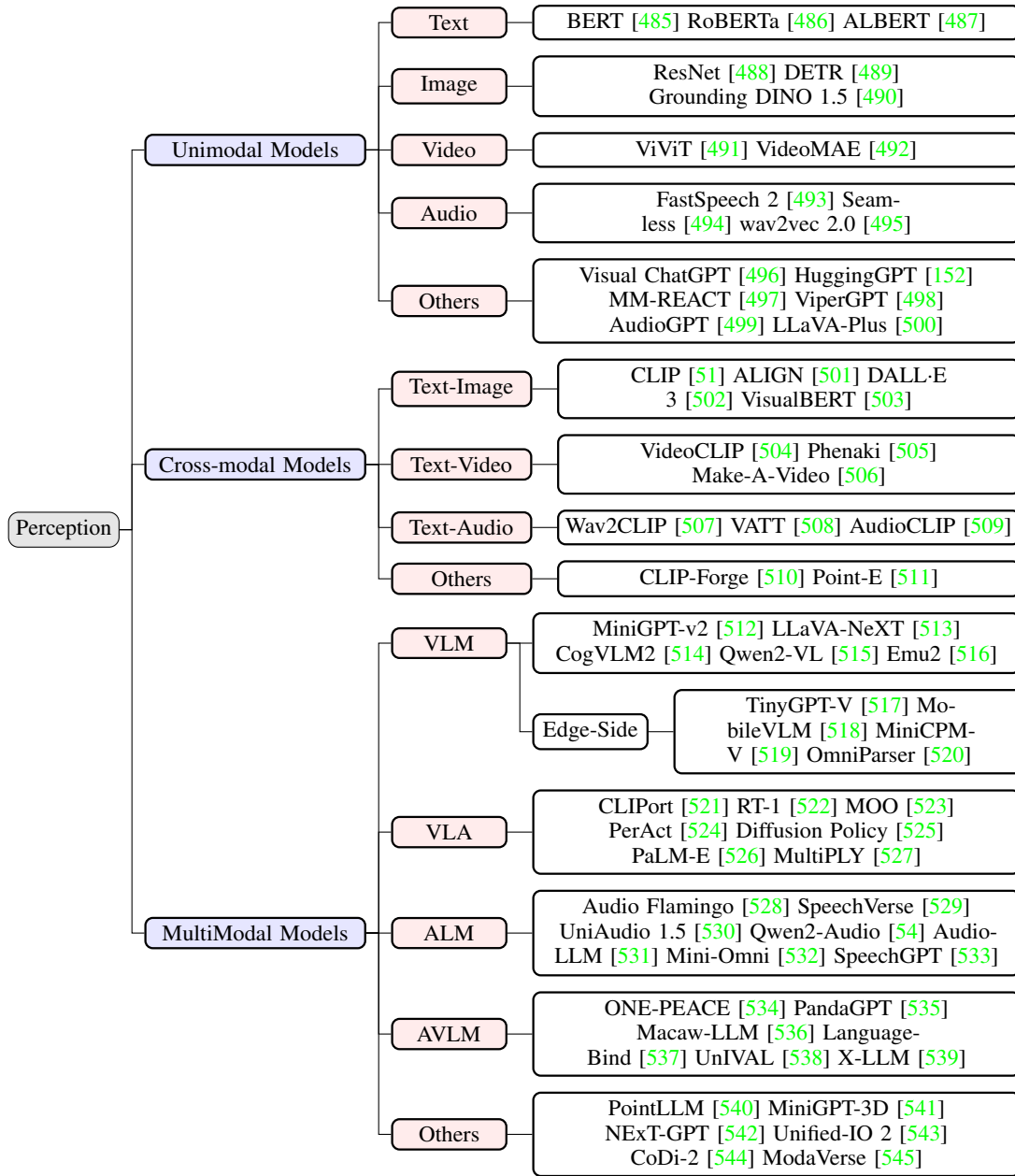


Figure 7.1: Illustrative Taxonomy of Perception System.

Another crucial distinction lies in perceptual processing efficiency. Human perception is limited by biological constraints such as nerve conduction speeds, typically in the range of milliseconds. Conversely, AI systems can process sensory inputs at speeds of microseconds or even nanoseconds, constrained primarily by computational hardware performance rather than biological limitations. Nevertheless, human perception naturally integrates information from multiple sensory modalities—known as multimodal perception—into coherent experiences effortlessly. For AI agents, achieving this multimodal integration requires carefully designed fusion algorithms that explicitly combine inputs from diverse sensors to build unified environmental representations [553].

Further differences arise in the way humans and artificial agents handle temporal and spatial information. Human perception is inherently continuous and fluid, smoothly experiencing the passage of time and spatial motion without explicit temporal discretization. In contrast, AI agents typically rely on discrete sampling of sensor data, using timestamps or sequential processing to simulate continuity. Spatial awareness in humans effortlessly merges visual, auditory, and vestibular information to achieve intuitive spatial positioning. For artificial agents, spatial perception usually involves

algorithmic processes such as simultaneous localization and mapping (SLAM) or 3D scene reconstruction from visual data sequences [554].

Physical or chemical stimuli transmitted from the external environment to human sensory organs will be received by the sensory system (such as eyes, ears, skin, etc.) and converted into neural signals, which are finally processed by the brain to produce perception of the environment. Similarly, to allow the intelligent agent to connect with the environment, it is also crucial to obtain these perception contents. Currently, various sensors are mainly used to convert electrical signals into processable digital signals. In this section, We distinguish between Unimodal models, Cross-modal models, and Multimodal models based on the number of modalities involved in the input and whether unified fusion modeling operations are performed. Unimodal Models specifically process and analyze data from a single modality or type of input (such as text, image, or audio), while Cross-modal Models establish relationships and enable translations between different modalities through dedicated mapping mechanisms, and Multimodal Models holistically integrate and process multiple modalities simultaneously to leverage complementary information for comprehensive understanding and decision-making.

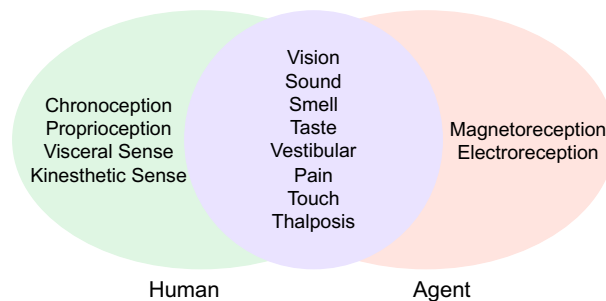


Figure 7.2: Comparison of common perceptual types between human and agent.

7.2 Types of Perception Representation

7.2.1 Unimodal Models

When humans are in an environment, they can listen to beautiful music, look at sunrise and sunset, or experience a wonderful audiovisual feast on stage. These perception contents can be either a single image or audio, or a fusion of multiple perception contents. Regarding the types of perception input of intelligent agents, we will start with single-modal and multimodal inputs, and introduce their implementation and differences.

Text As an important means of communication, text carries a wealth of information, thoughts, emotions and culture. Humans indirectly obtain the content of text through vision, hearing and touch, which is one of the most important ways for humans to interact with the environment. But for intelligent agents, text can directly serve as a bridge to connect with the environment, taking text as direct input and outputting response content. In addition to the literal meaning, text also contains rich semantic information and emotional color. In the early days, the bag-of-words model [555] was used to count text content and was widely used in text classification scenarios, but semantic expression could not be obtained. BERT [485] uses a bidirectional Transformer architecture for language modeling and captures the deep semantic information of text through large-scale unsupervised pre-training. [486, 487] further optimized the training efficiency of BERT. The autoregressive model represented by GPT3.5 [556] opened the prelude to LLM and further unified the tasks of text understanding and text generation, while technologies such as LoRA [109] greatly reduced the application cost of LLM and improved the agent’s perception ability of complex real-world scenario tasks.

Image Image is another important way for humans to interact with the environment which inherently encode spatial information, encompassing crucial attributes such as morphological characteristics, spatial positioning, dimensional relationships, and kinematic properties of objects. The evolution of computer vision architectures has demonstrated significant advancement in processing these spatial attributes. The seminal ResNet architecture [488] established foundational principles for deep visual feature extraction, while subsequent YOLO series [557, 558] demonstrated the capability to simultaneously determine object localization and classification with remarkable efficiency. A paradigm shift occurred with the introduction of DETR [489], which revolutionized object detection by implementing parallel prediction through global context reasoning, effectively eliminating traditional computational overhead associated with non-maximum suppression and anchor point generation. More recently, DINO 1.5 [490] has extended these capabilities to open-set scenarios through architectural innovations, enhanced backbone networks, and expanded training paradigms,

substantially improving open-set detection performance and advancing the perceptual generalization capabilities of artificial agents in unconstrained environments.

Video Video is an expression of continuous image frames, which includes the time dimension and displays dynamic information that changes over time through continuous image frames. The intelligent agent uses video as input and obtains richer perceptual content through continuous frames. ViViT [491] extracts spatiotemporal markers from videos, effectively decomposing the spatial and temporal dimensions of the input. VideoMAE [492] learns general video feature representations through self-supervised pre-training and has strong generalization capabilities on out-of-domain data. It lays a solid foundation for intelligent agents to acquire perceptual capabilities in new scenarios.

Audio In addition to text and vision, another important way for humans to interact with the environment is through audio. Audio not only contains direct text content, but also contains the speaker’s tone and emotion [559]. Wav2Vec2 [495] defines the contrast task by quantizing the potential representation of joint learning, achieving speech recognition effectiveness with 1/100 labeled data volume. FastSpeech 2 [493] directly introduces voice change information (pitch, energy, duration, etc.) and uses real targets to train the model to achieve more realistic text-to-speech conversion. Seamless [494] generates low-latency target translations through streaming and using an efficient monotonic multi-head attention mechanism, while maintaining the human voice style, to achieve synchronous speech-to-speech/text translation from multiple source languages to target languages. Based on these means, the intelligent agent can achieve the ability to listen and speak.

Others At present, most of the research on intelligent agents focuses on the above-mentioned common sensory input types. However, just as humans have more than 20 types of perception, intelligent agents have also made progress in achieving corresponding perception capabilities through other sensors. The bionic olfactory chip developed by Hong Kong University of Science and Technology [551] integrates a nanotube sensor array on a nanoporous substrate, with up to 10,000 independently addressable gas sensors on each chip, which is similar to the configuration of the olfactory system of humans and other animals, and can accurately distinguish between mixed gases and 24 different odors. In terms of taste, Tongji University [560] combines fluorescence and phosphorescence signals to develop an intelligent taste sensor with multi-mode light response, which can effectively identify umami, sourness and bitterness. In order to achieve human-like perception and grasping capabilities, New York University [561] launched a low-cost magnetic tactile sensor AnySkin, which can be quickly assembled and replaced. Even in the perception of pain, the Chinese Academy of Sciences uses the unique electrical properties of liquid metal particle films when they are “injured” (mechanically scratched) to imitate the perception and positioning of “wound.” Some other works, including HuggingGPT [152], LLaVA-Plus [500], and ViperGPT [498], integrate these single-modal perception capabilities within the framework, select and apply them according to task requirements, and achieve the goal of achieving more complex tasks.

7.2.2 Cross-modal Models

Text-Image Cross-modal models integrating text and images have witnessed significant advancements in recent years, leading to improved alignment, retrieval, and generation between the two modalities. These models can be categorized based on their primary objectives, including cross-modal alignment and retrieval, text-to-image generation, and image-to-text generation.

One of the primary focuses in cross-modal research is the alignment and retrieval of text and images. CLIP [51], introduced by OpenAI in 2021, employs contrastive learning to align textual and visual representations, enabling zero-shot cross-modal retrieval and classification. Similarly, ALIGN [501], developed by Google in the same year, leverages large-scale noisy web data to optimize text-image embedding alignment. In 2022, CyCLIP [562] introduced a cyclic consistency loss to further enhance the robustness of cross-modal alignment, improving the reliability of retrieval tasks.

Another major area of progress involves text-to-image generation, where models aim to synthesize high-quality images based on textual descriptions. OpenAI’s DALL-E series [563, 564, 502], spanning from 2021 to 2023, has made substantial contributions in this domain, with DALL-E 3 offering fine-grained semantic control over generated images. Stable Diffusion [565], introduced by Stability AI in 2022, employs a diffusion-based generative approach that supports open-domain text-to-image synthesis and cross-modal editing.

A third significant research direction is image-to-text generation, where models aim to generate high-quality textual descriptions based on image inputs. Typical representative work is the BLIP [566] and BLIP-2 [567] models, introduced by Salesforce between 2022 and 2023, which utilize lightweight bridging modules to enhance vision-language model integration, enabling tasks such as image captioning and question answering.

Text-Video The key research here involves video text alignment, generation and retrieval. VideoCLIP [504] employs a video encoder—typically based on temporal convolution or a transformer structure—to extract sequential features from video frames. These features are subsequently aligned with textual representations generated by a language encoder, facilitating robust video-text association. In the domain of text-to-video generation, Meta’s Make-A-Video model [506] extends spatial-temporal dimensions using diffusion-based techniques, allowing for high-quality video synthesis from textual descriptions. Additionally, Google’s Phenaki [505] addresses the challenge of generating long, temporally coherent video sequences, demonstrating significant advancements in video synthesis through cross-modal learning. DeepMind’s Frozen in Time [568] adopts contrastive learning for video-text matching, thereby enabling efficient cross-modal retrieval. This approach enhances the capacity to search and retrieve relevant video segments based on textual queries, further improving the integration of vision and language understanding.

Text-Audio Cross-modal models connecting text and audio have made significant improvements in related tasks such as modal representation, generation, and conversion, and enhanced the perception ability under a single modality.

AudioCLIP [509], introduced in 2021, extends the CLIP framework to the audio domain, enabling tri-modal retrieval across audio, text, and images. By incorporating audio as an additional modality, AudioCLIP utilizes multi-task learning to unify image, text, and audio representations into a shared embedding space. This advancement enhances the capability of cross-modal retrieval and interaction. In a similar vein, VATT [508] adopts a unified Transformer-based architecture to process video, audio, and text through independent encoding branches. These branches are subsequently fused into a shared multimodal space, facilitating tasks such as cross-modal retrieval and multi-task learning. This design allows for greater adaptability across diverse multimodal scenarios.

For text-to-audio generation, Meta introduced AudioGen [569] in 2023, which enables the synthesis of audio, such as environmental sounds and music fragments, directly from textual descriptions. This model exemplifies the growing capabilities of AI in generating high-fidelity audio based on linguistic input, expanding applications in media, entertainment, and accessibility.

Additionally, in the domain of speech-to-text and text-to-speech conversion, Microsoft developed SpeechT5 [570]. This model unifies speech and text generation, supporting both speech synthesis and recognition within a single framework. By leveraging a shared architecture for these dual functionalities, SpeechT5 contributes to the seamless integration of speech and text processing, thereby enhancing applications in automated transcription, voice assistants, and accessibility tools.

Others In some other scenarios and domains, cross-modal modeling also plays an important role.

CLIP-Forge [510] presents a novel method for generating 3D shapes from textual descriptions. By leveraging the capabilities of Contrastive Language-Image Pre-training (CLIP), this approach enables the synthesis of high-quality 3D objects conditioned on natural language inputs, bridging the gap between text and 3D geometry. Point-E [511] extends this concept by generating 3D point clouds from text descriptions. Unlike traditional 3D reconstruction techniques, Point-E focuses on point cloud representations, facilitating efficient and scalable 3D content creation while maintaining high fidelity to textual prompts.

In the field of medical imaging, MoCoCLIP [571] introduces an approach that enhances zero-shot learning capabilities. By integrating CLIP with Momentum Contrast (MoCo), this method improves the generalization of deep learning models in medical imaging applications, addressing the challenges associated with limited annotated data and domain adaptation.

7.2.3 Multimodal Models

The cross-modal model described above mainly aligns and maps between modalities through contrastive learning and other methods to achieve information complementarity and conversion between modalities. Furthermore, the work of multimodal models focuses on how to integrate the features of multiple data (such as vision, text, audio, etc.) to improve the performance of the overall model.

Vision Language Model Vision Language Model (VLM) is broadly defined as multimodal model that can learn from images (or videos) and text. Humans live in a world full of multimodal information. Visual information (such as images and videos) and language information (such as text) often need to be combined to fully express meaning. The same is true for intelligent agents. LLaVA [513] first tried to use gpt-4 to generate a multimodal language image instruction dataset. Through end-to-end training, a large multimodal model was obtained and excellent multimodal chat capabilities were demonstrated. LLaVA-NeXT [513] uses dynamic high-resolution and mixed data to show amazing zero-shot capabilities even in pure English modal data, and the computational/training data cost is 100-1000 times smaller than other methods. Emu2 [516] changes the traditional way of using image tokenizer to convert images into discrete tokens, and directly uses image encoders to convert images into continuous embeddings and provide them to Transformer,

enhancing multimodal context learning capabilities. MiniGPT-v2 [512] employs unique identifiers for various tasks during training. These identifiers help the model differentiate task instructions more effectively, enhancing its learning efficiency for each task. Qwen2-VL [515], DeepSeek-VL2 [572] use dynamic encoding strategies on visual components, aiming to process images with different resolutions and generate more efficient and accurate visual representations. At the same time, DeepSeek-VL2 [572] also uses the MoE model with a multi-head potential attention mechanism to compress the key-value cache into a latent vector to achieve efficient reasoning.

Previous work mainly uses image fusion text for training. Video-ChatGPT [573] extends the input to video and directly uses a video adaptive visual encoder combined with LLM for training to capture the temporal dynamics and inter-frame consistency relationships in video data, thereby enabling open conversations about video content in a coherent manner. To solve the lack of unified tokenization for images and videos, Video-LLaVA [574] unifies the visual representations of image and video encoding into the language feature space, making the two mutually reinforcing. Similarly, Chat-UniVi [575] employs a set of dynamic visual tokens to integrate images and videos, while utilizing multi-scale representations to allow the model to grasp both high-level semantic concepts and low-level visual details. Youku-mPLUG [576] has made in-depth research in specific scenarios. Based on the high-quality Chinese video-text pairs in the Youku video sharing platform, it enhances the ability to understand overall and detailed visual semantics and recognize scene text. Unlike the previous method that requires training, SlowFast-LLaVA [577] can effectively capture the detailed spatial semantics and long-term temporal context in the video through a two-stream SlowFast design without any additional fine-tuning of the video data, achieving the same or even better results than the fine-tuning method.

As the parameters of large models gradually decrease and the computing power of the end-side increases, high-performance end-side models are gaining momentum. Smart terminal devices such as mobile phones and PCs have strong demands for image visual processing, which puts forward higher multimodal recognition effects and reasoning performance requirements for the deployment of AI models on the end-side. TinyGPT-V [517] is built based on the Phi-2 [578] small backbone combined with BLIP-2 [567], only 8G video memory or CPU is needed for reasoning, and solving the computational efficiency problems of LLaVA [513] and MiniGPT-4 [579]. MiniCPM-V [519] mainly provides powerful OCR capabilities for long and difficult images, and has a low hallucination rate, providing reliable perception output. Megrez-3B-Omni [580] ensures that all structural parameters are highly compatible with mainstream hardware through coordinated optimization of software and hardware. Its inference speed is up to 300% faster than that of models with the same precision, improving its adaptability to different end-side hardware.

Similarly, there are more GUI-related works focusing on automatic task execution on mobile phones and PCs. Omni-Parser [520] uses popular web page and icon description datasets for fine-tuning, significantly enhancing the detection and functional semantic expression capabilities of icons in screenshots. GUICourse [581] and OS-ATLAS [582] also built a cross-platform GUI grounding corpus, which brought significant performance improvements in the understanding of GUI screenshots and enriching the interactive knowledge of GUI components.

Vision Language Action Model Vision-Language-Action (VLA) model, which takes vision and language as inputs and generates robotic actions as outputs, represents an important research direction in the field of embodied intelligence. The selection of vision and language encoders in VLA models has undergone diverse development, evolving from early CNNs to Transformer architectures, and further integrating 3D vision and large language models. Early models such as CLIPort [521] used ResNet [488] to process visual inputs and combined language embeddings to generate actions, laying the foundation for multimodal fusion. RT-1 [522] introduced the Transformer architecture, employing EfficientNet as the visual encoder and USE as the language encoder, and fused visual and language information via FiLM mechanisms, significantly enhancing the model’s generalization ability. VIMA [523] further adopted multimodal prompts, combining the ViT visual encoder and the T5 language model to support more complex tasks. PerAct [524] innovatively used 3D point clouds as visual inputs and processed multi-view information through Perceiver IO, providing richer spatial perception for robotic manipulation. Diffusion Policy [525] combined ResNet visual encoders and Transformer language models, generating actions through diffusion models to improve the diversity and accuracy of action generation. SayCan [583] integrated the PaLM language model with visual inputs, using the CLIP visual encoder for task decomposition. PaLM-E [526] combined the ViT visual encoder and the PaLM language model, guiding low-level action execution through text planning. MultiPLY [527] further integrated 3D information into LLMs, combining the EVA visual encoder and the LLaMA language model to provide more comprehensive planning capabilities for complex tasks.

Audio Language Model Audio Language Model(ALM) uses the audio and text to build multimodal model. Speechgpt [533] built a large-scale cross-modal speech instruction dataset SpeechInstruct and trained discrete speech representations, achieving cross-modal speech dialogue capabilities beyond expectations. LauraGPT [584], unlike the previous sampling of discrete audio tokens to represent input and output audio, proposed a novel data representation that combines the continuous and discrete features of audio, and demonstrated excellent performance on a wide range of

audio tasks through supervised multi-task learning. [529, 585, 531] converts audio data into embedded representations and then fine-tunes instructions, so that excellent performance can be achieved on various speech processing tasks through natural language instructions. In order to reduce the cost of fine-tuning training, Audio Flamingo [528] quickly enhances the ability to adapt to unseen tasks through contextual learning and retrieval based on the audio language model. UniAudio 1.5 [530] uses words or subwords in the text vocabulary as audio tokens, learns these audio representations through a small number of samples, and achieves cross-modal output without fine-tuning. In order to make the output more realistic and in line with human expectations, Qwen2-Audio [54] introduced the DPO training method to achieve human preference alignment.

Audio Vision Language Model Audio Vision Language Model (AVLM) utilizes audio, vision, and text to unify multimodal models. Previously, we introduced some work on building multimodal models using information from two modalities. In the pursuit of AGI, the obstacle to achieving this goal lies in the diversity and heterogeneity of tasks and modalities. A suitable approach is to allow more modal capabilities to be supported within a unified framework. Some closed-source work [586, 587] has achieved excellent capabilities across modalities such as text, vision, and audio. ImageBind [588] implements joint embedding across six different modes (image, text, audio, depth, thermal, and IMU data). Panda-GPT [535] combines ImageBind’s multi-modal encoder and Vicuna [589], showing zero-shot cross-modal performance in addition to images and text. Similar work includes [539, 539, 536], which achieves alignment and training through the encoding information of vision, audio and text. Multimodal models often require more resources to train, and UniVAL [538] trained a model with only $\sim 0.25B$ parameters based on task balance and multimodal curriculum learning, and used weight interpolation to merge multimodal models, maintaining generalization under out-of-distribution. NExT-GPT [542] connects LLM with multimodal adapters and different diffusion decoders, and only trains a small number of parameters (1%) of certain projection layers.

Other works [543, 590, 544, 545] have achieved input-output conversion between arbitrary modalities. Unified-IO 2 [543] is the first autoregressive multimodal model that can understand and generate images, text, audio, and actions. It tokenizes different modal inputs into a shared semantic space and processes them using an encoder-decoder model. AnyGPT [590] builds the first large-scale any-to-any multimodal instruction dataset, using discrete representations to uniformly process various modal inputs. Modaverse [545] directly aligns the output of the LLM with the input of the generative model to solve the problem that previous work relies heavily on the alignment of the latent space of text and non-text features, avoiding the complexity associated with the alignment of latent features. CoDi-2 [544] outperforms earlier domain-specific models in tasks like topic-based image generation, visual transformation, and audio editing.

Others Humans have explored the 2D world more than the 3D world, but 3D can more accurately describe the shape and texture information of objects and provide richer perceptual information. PointLLM [540] uses a point cloud encoder to express geometric and appearance features, and integrates language features for two-stage training of complex point-text instructions, achieving excellent 3D object description and classification capabilities. Since 3D contains richer information than 2D, it also brings greater training costs. [541, 591] reduces the training cost here, and MiniGPT-3D [541] uses 2D priors from 2D-LLM to align 3D point clouds with LLMs. Modal alignment is performed in a cascade manner, and query expert modules are mixed to efficiently and adaptively aggregate features, achieving efficient training with small parameter updates. LLaVA-3D [591] connects 2D CLIP patch features with their corresponding positions in 3D space, integrates 3D Patches into 2D LMM and uses joint 2D and 3D visual language command adjustment to achieve a 3.5-fold acceleration in convergence speed.

In order to enable intelligent agents to accurately perceive and manipulate unknown objects, Meta [592] developed NeuralFeels technology, which combines vision and touch to continuously model unknown objects in 3D, more accurately estimate the posture and shape of objects in handheld operations, and improve the accuracy of ignorant object operations by 94%.

7.3 Optimizing Perception Systems

Perception errors, including inaccuracies, misinterpretations, and “hallucinations” (generation of false information), pose substantial challenges to the reliability and effectiveness of LLM-based agents. Optimizing perception thus requires minimizing these errors using various strategies across model, system, and external levels.

7.3.1 Model-Level Enhancements

Fine-tuning. Fine-tuning pre-trained LLMs on domain-specific data significantly improves their ability to accurately perceive and interpret relevant information. For example, fine-tuning models such as LLaVA on specific landmarks has been shown to enhance their recognition accuracy, particularly in urban navigation tasks [513, 593]. Moreover, techniques such as Low-Rank Adaptation (LoRA) enable more efficient fine-tuning, avoiding a substantial increase in

model complexity while still improving performance [109, 594]. Some LLM work combined with traditional vision is also widely used. Integrating with YOLOs [595] on the basis of the the Llama-Adapter [596] architecture significantly improves the detection and positioning capability.

Prompt Engineering. The design of effective prompts is crucial to ensure LLMs generate outputs that are both accurate and aligned with the desired goals. By providing clear instructions, contextual information, and specific formatting requirements, prompt engineering minimizes misinterpretation and hallucination [597]. System prompts define the agent’s role, historical prompts to provide context from past interactions, and customized prompts to ensure output consistency has been shown to reduce errors significantly [597].

Retrieval-Augmented Generation. Supplementing LLMs with external knowledge sources through retrieval mechanisms helps ground their responses in factual information, reducing the likelihood of hallucinations and improving the accuracy of perceived information [334].

7.3.2 System-Level Optimizations

Anticipation-Reevaluation Mechanism. In scenarios where agents face incomplete or ambiguous information, an anticipation-reevaluation mechanism can enhance robustness. For instance, in navigation tasks, agents can anticipate goal directions based on historical data and reevaluate their inferences when new information becomes available [598].

Multi-Agent Collaboration. In multi-agent systems, structured communication and collaboration among agents can facilitate information sharing, error correction, and consensus-building, leading to a more accurate collective perception of the environment [599]. Different communication topologies, such as fully connected, centralized, and hierarchical structures, offer varying trade-offs in terms of efficiency and robustness [600]. InsightSee [601] refines visual information through a multi-agent framework with description, reasoning, and decision-making, effectively enhancing visual information processing capabilities. Similarly, HEV [602] integrates the global perspective information of multiple agents and endows RL agents with global reasoning capabilities through cooperative perception, thereby enhancing their decision-making capabilities.

Agent Specialization. Assigning distinct roles and capabilities to individual agents within a multi-agent system allows for a division of labor in perception, with each agent focusing on specific aspects of the environment or task. This can enhance the overall accuracy and efficiency of perception [603].

7.3.3 External Feedback and Control

Loss Agents for Optimization. Utilizing LLMs as loss agents, allows for the dynamic adjustment of loss function weights during training [604]. This enables the optimization of image processing models based on complex, potentially non-differentiable objectives, including human feedback and evaluations from specialized models. This approach essentially externalizes the optimization objective, allowing the LLM to “perceive” and adapt to complex criteria [605].

Human-in-the-Loop Systems. Incorporating human feedback and oversight can help correct errors, guide the agent’s learning process, and ensure alignment with human values and expectations [43].

Content and Output Mediation. Before presenting LLM outputs to users, content mediation filters and refines these outputs. This helps prevent unexpected or harmful behaviors, ensuring alignment with user expectations and safety guidelines [606].

7.4 Perception Applications

The operational efficacy of intelligent agents is predominantly influenced by three critical factors: model architecture dimensionality, hardware infrastructure specifications, and quantization optimization methodologies. The exponential progression in model parameters—from Bert-Base’s modest 110M to GPT-3’s substantial 175 billion, culminating in Llama 3’s unprecedented 405 billion—has correspondingly escalated processing latency from milliseconds to hundreds of milliseconds. Hardware performance variations are particularly noteworthy; empirical evidence with GPT-3 demonstrates that NVIDIA H100 exhibits a 50% improvement in token processing throughput compared to A100, while RTX 4090 achieves approximately double the processing capability.

Contemporary intelligent agents have penetrated diverse domains, encompassing personal assistance systems, gaming environments, Robotic Process Automation (RPA), and multimedia content generation, predominantly leveraging visual perception as their primary input modality. In the context of procedurally generated environments like Minecraft, STEVE [607] demonstrates remarkable performance improvements, achieving a 1.5x acceleration in technology tree progression and a 2.5x enhancement in block search efficiency through visual information processing. Steve-Eye [608]

advances this paradigm through end-to-end multimodal training, addressing environmental comprehension latency through integrated visual-textual input processing.

In creative content generation, AssistEditor [609] exemplifies sophisticated multi-agent collaboration, facilitating professional video editing through style-driven content understanding. Similarly, Audio-Agent [610] implements cross-modal integration between textual/visual inputs and audio outputs, enabling comprehensive audio manipulation capabilities [611, 612, 613].

Mobile and desktop platforms have witnessed significant advancements in agent applications. ExACT [614] has established new state-of-the-art benchmarks in VisualWebArena [615], achieving a 33.7% Success Rate through screenshot-based exploratory learning with caption and Set of Mask integration. SPA-Bench [616] introduces a comprehensive mobile evaluation framework that authentically replicates real-world complexity. M3A [617] demonstrates superior performance with a 64.0% success rate in SPA-Bench through multimodal input processing. AgentStore [618] has markedly improved OSWorld PC benchmark performance to 23.85% through enhanced visual and accessibility tree processing.

Voice interaction capabilities [619, 586] in personal AI assistants have significantly reduced interaction friction while enhancing operational efficiency. The integration of emotional prosody in voice interactions has demonstrated increased user engagement and retention.

In embodied intelligence applications, haptic and force feedback mechanisms have emerged as crucial modalities for environmental interaction, with enhanced sensory fidelity enabling increasingly precise operational capabilities [620].

7.5 Summary and Discussion

Although more and more research works [543, 590] focus on building unified multimodal models to support the input and output of multiple perception capabilities. Agent perception, a cornerstone of autonomous systems, faces significant challenges in effectively interpreting and integrating multi-modal data. Current methodologies encounter persistent issues in representation learning, alignment, and fusion, which hinder the development of robust and generalizable perception systems.

One of the primary issues lies in the representation methods employed, which often fail to capture the intricate nuances of multi-modal data. This shortfall is particularly evident in scenarios where high-dimensional sensory inputs require a sophisticated abstraction that preserves critical semantic information. Furthermore, the alignment of representations presents additional difficulties. Integrating heterogeneous data types into a cohesive feature space is not only computationally intensive but also prone to inconsistencies, which can lead to misinterpretation of ambiguous signals. The challenge is compounded when attempting to fuse these diverse representations, as the process of merging features from various sources frequently results in suboptimal integration and potential loss of vital information.

Future research directions should prioritize adaptive representation learning through dynamic neural architectures capable of automatically adjusting their structure based on environmental context and task demands. This could involve meta-learned parameterization or graph-based representations that explicitly model relationships between perceptual entities. For cross-modal alignment, self-supervised spacetime synchronization mechanisms leveraging contrastive learning principles show promise in establishing dense correspondence without requiring exhaustive labeled data. The integration of causal inference frameworks into alignment processes [621] could further enhance robustness against spurious correlations. In representation fusion, hierarchical attention mechanisms with learnable gating functions merit deeper exploration to enable context-aware integration of complementary modality features. Emerging techniques in differentiable memory networks may provide new pathways for maintaining and updating fused representations over extended temporal horizons.